AD-A068 551 MASSACHUSETTS INST OF TECH LEXINGTON LINCOLN LAB F/G 12/1
ON OPTIMAL DESCRIMINANTS BETWEEN TWO CLASSES OF RANDOM VARIABLE--ETC(U)
FEB 79 L K JONES F19628-78-C-0002
UNCLASSIFIED TN-1979-22 ESD-TR-79-31 NL

| OF |

AD
A068551

END
DATE
FILMED
6-79
DDC

[illegible] Note

1979-23

S. K. Jones

[illegible]

28 February 1979

[illegible]

Lincoln Laboratory

[illegible]

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

LINCOLN LABORATORY

ON OPTIMAL DISCRIMINANTS BETWEEN TWO CLASSES OF RANDOM VARIABLES IN TERMS OF THE MOMENTS OF THEIR DISTRIBUTIONS

L. K. JONES

Group 92

TECHNICAL NOTE 1979-22

28 FEBRUARY 1979

LEXINGTON MASSACHUSETTS

ABSTRACT

For many problems of interest in statistical pattern recognition, density estimates for a random variable X of dimension d are unreliable unless the number of sample vectors is very large ($>2^d$). For even moderately large d ($d > 12$), sample sizes are often insufficient. However, lower order moments of the form $x_i^q \; x_j^\ell$ may be accurately estimated. In this paper we are concerned with the problem of optimally discriminating between two classes of random variables in terms of the available information about them of reasonable accuracy (their lower order moments). In no case do we make any assumption about the form of the probability densities of random variables X. (We do in some cases assume certain forms for the densities of functions of these random variables L(X).)

First we consider the performance of the Gaussian discriminant function for arbitrary class distributions. How well this function discriminates depends on the magnitudes of the interclass differences of the first four moments. It is shown how to adjust the constant term of the Gaussian discriminant to obtain minimum probability of error. Then a second order solution for the optimal quadratic discriminant is given. Finally the methods developed are applied to determine general discriminant functions.

CONTENTS

I. INTRODUCTION

Suppose two classes of events are mutually exclusive and exhaustive. Let these classes be denoted by ω_1 and ω_2 and their respective probabilities by $P(\omega_1)$ and $P(\omega_2)$. Suppose further that each event has associated with it a vector X in R^d. This vector contains the only observable information about the event. We then have a random variable defined on each class. Denote the probability densities of these random variables by $p_1(X)$ and $p_2(X)$. Let M_1, M_2, Σ_1 and Σ_2 be the means and covariance matrices for these densities. Although we may indeed not know the form of the densities $p_1(X)$ and $p_2(X)$, we may estimate (or know) the first and second moments listed above as well as various higher moments.

For many practical problems this is the case. A collection of sample X's of class ω_1 and sample X's of class ω_2 are used to estimate the densities $p_1(X)$, $p_2(X)$ and the moments. However small sample *sizes* ($<2^d$) *prevent us from* accurately estimating these densities. But moments of the form $x_i^q \, x_j^\ell$ may be reliably estimated from samples of sizes which are not exponential in d. Hence in many actual situations, where d is only moderately large ($d > 12$), the only reliable information about the classes ω_1 and ω_2 consists of lower order moments. Using this information how do we best discriminate between ω_1 and ω_2? This is exactly the mathematical problem we shall consider.

If we observe a certain vector X, how do we decide with which class it is most likely associated? One method is the Gaussian discriminant defined as follows: Assign an observed test vector, X, to class 1 or 2 according to the magnitude of the following expression -

$$\tfrac{1}{2}(X-M_2)^t \Sigma_2^{-1}(X-M_2) - \tfrac{1}{2}(X-M_1)^t \Sigma_1^{-1}(X-M_1) + \tfrac{1}{2} \ln |\Sigma_2| - \tfrac{1}{2} \ln |\Sigma_1|$$

$$\begin{cases} > t \text{ class } 1 \\ \leq t \text{ class } 2 \end{cases}$$

where the threshold

$$t = \ln P(\omega_1) - \ln P(\omega_2)$$

If $p_1(X)$ and $p_2(X)$ are normal densities this is equivalent to choosing class 1 if $\frac{p_1(X)}{p_2(X)} > \frac{P(\omega_1)}{P(\omega_2)}$ and class 2 otherwise.

Clearly this procedure is optimal (minimizes the probability of error) in the normal case.

Let us call the above discriminant function $G(X)$. Although G may be the optimal discriminant function only in the case that p_1 and p_2 are indeed normal, it may be a good approximation to an optimal discriminant in other cases. We shall first prove a result in this direction and then show how to adjust the constant term of the discriminant $G(X)$ to yield minimum probability of error using third and fourth moments. We then give minimal variance[1] solutions to the problem of finding the optimal quadratic discriminant function. These solutions are of minimum error in several asymptotic cases. Again the first four moments are used. Finally general discriminant functions are derived using marginal distributions or moments of the form $x_i^q x_j^\ell$. Upper bounds on the probability of error of these discriminants are given in asymptotic cases using the

[1] These solutions, introduced in section III, are so named since they involve the minimization of certain weighted sums of the variances of a discriminant function under each hypothesis.

central limit theorem for finitely dependent random variables (see [3]).

II. THE GAUSSIAN DISCRIMINANT

Theorem 1[1] $E_{p_1}(G) = \int_{R^d} G(X)p_1(X)dX \geq 0$

$$\geq \int_{R^d} G(X)p_2(X)dX = E_{p_2}(G)$$

Proof: Suppose $Y = A\,X$ is a non-singular linear mapping. For a probability density δ we have $M_{\tilde\delta} = A\,M_\delta$ and $\Sigma_{\tilde\delta} = A\,\Sigma_\delta\,A^t$ where $M_{\tilde\delta}$ and $\Sigma_{\tilde\delta}$ are the mean and covariance matrix for the transformed density $\tilde\delta(Y)=|A|^{-1}\delta(A^{-1}Y)$. As on page 34 of [1] let us determine A such that

$$\Sigma_{\tilde 1} = A\,\Sigma_1\,A^t = \begin{pmatrix} 1 & & & 0 \\ & 1 & & \\ & & \ddots & \\ 0 & & & 1 \end{pmatrix} = I$$

$$\Sigma_{\tilde 2} = A\,\Sigma_2\,A^t = \begin{pmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_d \end{pmatrix} =$$

where $\tilde i$ denotes $\tilde p_i$.

[1] This result was stated without proof in [2] for the univariate case.

Now $E_{p_1}\ (G(X)) = E_{\tilde{p}_1}\ \Big(G(X(Y))\Big) =$

$$E_{\tilde{p}_1}\left[\frac{1}{2}(Y-M_{\tilde{2}})^t\Lambda^{-1}(Y-M_{\tilde{2}}) - \frac{1}{2}(Y-M_{\tilde{1}})^t(Y-M_{\tilde{1}}) + \frac{1}{2}\ln|A\,\Sigma_2A^t| - \frac{1}{2}\ln|A\,\Sigma_1A^t|\right]$$

$$= E_{\tilde{p}_1}\left[\frac{1}{2}\sum_1^d\left[\frac{(y_i-M_{\tilde{2}i})^2}{\lambda_i} - (y_i - M_{\tilde{1}i})^2 + \ln\lambda_i\right]\right]$$

$$= \frac{1}{2}\sum_1^d\left[\frac{1}{\lambda_i}E_{\tilde{p}_1}\Big[(y_i-M_{\tilde{1}i}) + (M_{\tilde{1}i}-M_{\tilde{2}i})\Big]^2 - 1 + \ln\lambda_i\right]$$

$$= \frac{1}{2}\sum_1^d\left[\frac{1}{\lambda_i} + \frac{1}{\lambda_i}(M_{\tilde{1}i}-M_{\tilde{2}i})^2 - 1 + \ln\lambda_i\right] \geq 0$$

since $0 < \lambda_i < \infty$ and $z-1 \geq \ln z$ for all $z > 0$.

Similarly $E_{p_2}\ [Q(X)] = E_{\tilde{p}_2}\ [Q(X(Y))]$

$$= \frac{1}{2}\sum_1^d\left[1 - E_{\tilde{p}_2}\Big[(y_i-M_{\tilde{2}i}) + (M_{\tilde{2}i}-M_{\tilde{1}i})\Big]^2 + \ln\lambda_i\right]$$

$$= \frac{1}{2}\sum_1^d\Big(1 - \lambda_i - (M_{\tilde{2}i}-M_{\tilde{1}i})^2 + \ln\lambda_i\Big)$$

$$\leq \frac{1}{2}\sum_1^d\Big(1-\lambda_i + \ln\lambda_i\Big) \leq 0$$

This completes the proof of the theorem.

From the above $E_{p_1}(G(X)) - E_{p_2}(G(X)) =$

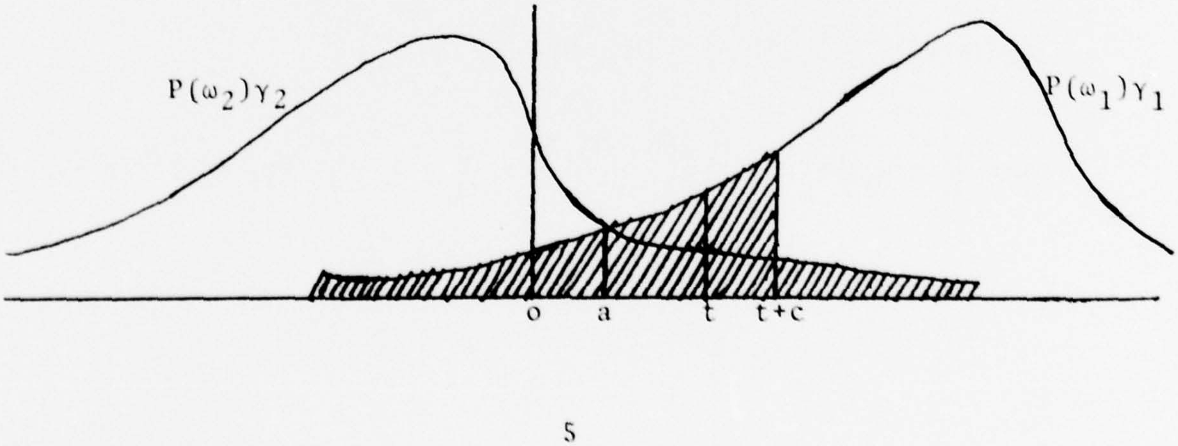
$$\frac{1}{2}\sum_{1}^{d}\left[(\tilde{M}_{2i}-\tilde{M}_{1i})^{2}\;(1+\frac{1}{\lambda_i}) + \frac{1}{\lambda_i} + \lambda_i - 2\right]$$ which is the

divergence when p_1 and p_2 are normal densities. This quantity is always positive when the first two moments of $(X|\omega_1)$ and $(X|\omega_2)$ are not identical. This follows since $\frac{1}{z} + z > 2$ for all $z > 0$, $z \neq 1$.

We now discuss the adjustment of the constant term in $G(X)$. This is equivalent to finding c which minimizes the error of the decision rule $G - c \gtrless t$. This error function may be written

$$E(c) = P(\omega_2)\int_{G-c>t} dp_2 + P(\omega_1)\int_{G-c\leq t} dp_1$$

If p_1 and p_2 are normal densities, then c=0 is optimal. For the general case consider the graphs of $P(\omega_1)\gamma_1$ and $P(\omega_2)\gamma_2$, where γ_1 and γ_2 are the density functions of the random variables $G(X/\omega_1)$ and $G(X/\omega_2)$ respectively.



E(c) will then be the area of the shaded region. This is minimum for c = a-t where a is a solution of the equation $P(\omega_2)\gamma_2$ (a) = $P(\omega_1)\gamma_1$ (a). Again in the normal case a=t by the optimality of the decision rule G $\gtrless$ t. This is however not always true in the general case as will be demonstrated below.

We now estimate a. Since both $G(X/\omega_1)$ and $G(X/\omega_2)$ can be expressed as sums of d random variables, the central limit theorem implies in many cases that γ_1 and γ_2 approach normal densities for large d. This occurs for instance when the y_i in the proof of the above theorem are independent under the hypotheses ω_1 and ω_2. We then estimate a by a solution of

$$\frac{\left[a - E(G(X/\omega_1))\right]^2}{2\,\mathrm{Var}(G(X/\omega_1))} = \frac{\left[a - E(G(X/\omega_2))\right]^2}{2\,\mathrm{Var}(G(X/\omega_2))} + \ln\left(\frac{P(\omega_1)}{P(\omega_2)}\sqrt{\frac{\mathrm{Var}(G(X/\omega_2))}{\mathrm{Var}(G(X/\omega_1))}}\right)$$

The calculations are as follows:

$$E(G(X/\omega_1)) = \frac{1}{2}\sum_{1}^{d}\left(\frac{1}{\lambda_i} + \frac{1}{\lambda_i}(M_{\tilde{1}i} - M_{\tilde{2}i})^2 - 1 + \ln \lambda_i\right)$$

$$E(G(X/\omega_2)) = \frac{1}{2}\sum_{1}^{d}\left(1 - \lambda_i - (M_{\tilde{2}i} - M_{\tilde{1}i})^2 + \ln \lambda_i\right)$$

$$\mathrm{Var}\left(G(X/\omega_1)\right) = E_{\tilde{p}_1}\left[\frac{1}{2}\sum_{1}^{d}\left(\frac{(y_i - M_{\tilde{2}i})^2}{\lambda_i} - \frac{1}{\lambda_i} + 1 - \frac{1}{\lambda_i}(M_{\tilde{1}i} - M_{\tilde{2}i})^2 - (y_i - M_{\tilde{1}i})^2\right)\right]^2$$

$$= E_{\tilde{p}_1} \left[\frac{1}{2} \sum_{1}^{d} \left((y_i - M_{\tilde{1}i})^2 (\frac{1}{\lambda_i} - 1) + \frac{2(y_i - M_{\tilde{1}i})(M_{\tilde{1}i} - M_{\tilde{2}i})}{\lambda_i} - (\frac{1}{\lambda_i} - 1) \right) \right]^2$$

$$= E_{\tilde{p}_1} \left[\frac{1}{4} \sum_{i,j=1}^{d} \left((y_i - M_{\tilde{1}i})^2 (y_j - M_{\tilde{1}j})^2 (\frac{1}{\lambda_i} - 1)(\frac{1}{\lambda_j} - 1) + (\frac{1}{\lambda_i} - 1)(\frac{1}{\lambda_j} - 1) - 2(y_i - M_{\tilde{1}i})^2 (\frac{1}{\lambda_i} - 1)(\frac{1}{\lambda_j} - 1) \right.\right.$$

$$\left.\left. - \frac{4(y_i - M_{\tilde{1}i})^2 (M_{\tilde{1}j} - M_{\tilde{2}j})(\frac{1}{\lambda_i} - 1)(y_j - M_{\tilde{1}j})}{\lambda_j} + \frac{4(y_i - M_{\tilde{1}i})(y_j - M_{\tilde{1}j})(M_{\tilde{1}i} - M_{\tilde{2}i})(M_{\tilde{1}j} - M_{\tilde{2}j})}{\lambda_i \lambda_j} \right) \right]$$

$$= \frac{1}{4} \sum_{i,j=1}^{d} (\frac{1}{\lambda_i} - 1)(\frac{1}{\lambda_j} - 1) \; E_{\tilde{p}_1} \left[(y_i - M_{\tilde{1}i})^2 (y_j - M_{\tilde{1}j})^2 - 1 \right]$$

$$+ \sum_{i,j=1}^{d} \frac{(\frac{1}{\lambda_i} - 1)(M_{\tilde{1}i} - M_{\tilde{2}i})}{\lambda_j} \; E_{\tilde{p}_1} \left[(y_i - M_{\tilde{1}i})^2 (y_j - M_{\tilde{1}j}) \right]$$

$$+ \sum_{i=1}^{d} \frac{(M_{\tilde{1}i} - M_{\tilde{2}i})^2}{\lambda_i^2}$$

$$\mathrm{Var} \left(G(X/\omega_2) \right) = E_{\tilde{p}_2} \left[\frac{1}{2} \sum_{1}^{d} \left(\frac{(y_i - M_{\tilde{2}i})^2}{\lambda_i} - (y_i - M_{\tilde{1}i})^2 - 1 + \lambda_i + (M_{\tilde{2}i} - M_{\tilde{1}i})^2 \right) \right]^2$$

$$= E_{\tilde{p}_2} \left[\frac{1}{2} \sum_{1}^{d} (y_i - M_{\tilde{2}i})^2 (\frac{1}{\lambda_i} - 1) - 2(y_{\tilde{i}} - M_{\tilde{2}i})(M_{\tilde{2}i} - M_{\tilde{1}i}) - (1 - \lambda_i) \right]^2$$

$$\begin{aligned}
= E_{p_2} \Bigg[& \frac{1}{4} \sum_{i,j=1}^{d} (y_i - M_{\tilde{2}i})^2 (y_j - M_{\tilde{2}j})^2 \left(\frac{1}{\lambda_i} - 1\right)\left(\frac{1}{\lambda_j} - 1\right) + \\
& (1-\lambda_i)(1-\lambda_j) \\
& -4 (y_i - M_{\tilde{2}i})^2 (y_j - M_{\tilde{2}j}) (M_{\tilde{2}j} - M_{\tilde{1}j}) \left(\frac{1}{\lambda_i} - 1\right) \\
& -2 (y_i - M_{\tilde{2}i})^2 \left(\frac{1}{\lambda_i} - 1\right)(1-\lambda_j) \\
& +4 (y_i - M_{\tilde{2}i})(y_j - M_{\tilde{2}j})(M_{\tilde{2}i} - M_{\tilde{1}i})(M_{\tilde{2}j} - M_{\tilde{1}j}) \Bigg] \\
= & \frac{1}{4} \sum_{i,j=1}^{d} \left(\frac{1}{\lambda_i} - 1\right)\left(\frac{1}{\lambda_j} - 1\right) E_{\tilde{p}_2} \left[(y_i - M_{\tilde{2}i})^2 (y_j - M_{\tilde{2}j})^2 - \lambda_i \lambda_j \right] \\
& + \sum_{i,j=1}^{d} \left(\frac{1}{\lambda_i} - 1\right)(M_{\tilde{1}j} - M_{\tilde{2}j}) E_{\tilde{p}_2} \left[(y_i - M_{\tilde{2}i})^2 (y_j - M_{\tilde{2}j}) \right] \\
& + \sum_{i=1}^{d} \lambda_i (M_{\tilde{1}i} - M_{\tilde{2}i})^2
\end{aligned}$$

The above third and fourth moments may be easily estimated from sample data by using the transformation A of the theorem. The solution we then choose for a is that solution for which $P(\omega_1)\gamma_1$ is increasing more rapidly than $P(\omega_2)\gamma_2$, i.e., for which

$$\frac{E(G(X/\omega_1))-a}{\mathrm{Var}(G(X/\omega_1))} \geq \frac{E(G(X/\omega_2))-a}{\mathrm{Var}(G(X/\omega_2))}.$$

III. THE MINIMAL VARIANCE SOLUTION

Def. 1: Let D be a class of probability densities $\mu(m,v)$ on the real line parametrized continuously by their means m and variances v $(m \varepsilon R,\ v \varepsilon V \subset R)$. D is called translational if in addition

$$\mu(m_1,v)(x) = \mu(m_2,v)(x+m_2-m_1) \text{ for all real } x.$$

Def. 2: Let D be a translational family. Then D is said to be of monotone error if in addition to the above the following is satisfied: For each $0 < \alpha < 1$ and any two members of D, $\mu(0,v_1)$ and $\mu(1,v_2)$, the function

$$E_\alpha(v_1,v_2) = \min_c \left\{ \int_{-\infty}^{c} (1-\alpha)\mu(1,v_2)dx + \int_{c}^{+\infty} \alpha\, \mu(0,v_1)dx \right\}$$

is differentiable in v_1 and v_2 and

$$\left|\frac{\partial E_\alpha}{\partial v_2}\right| + \left|\frac{\partial E_\alpha}{\partial v_1}\right| > 0 \text{ for all } v_1,\ v_2 \in V.$$

Theorem 2 Let $v_1(\vec{a})$, $v_2(\vec{a})$ be differentiable mappings from some parameter space $A \subset R^n$ into the variance space V of a monotone error family D. If $\vec{a}'$ is a local extremum of the function $E_\alpha\ (v_1(\vec{a}),v_2(\vec{a}))$ then $\vec{a}'$ is a local extremum of $\beta v_1(\vec{a}) + (1-|\beta|)v_2(\vec{a})$ for some $-1 \leq \beta \leq 1$.

proof: Let $\vec{a}'$ be a local extremum of $E_\alpha(v_1(\vec{a}),v_2(\vec{a}))$.

Then $\nabla E_\alpha(v_1(\vec{a}'),v_2(\vec{a}')) =$

$$\left(\frac{\partial E_\alpha}{\partial v_1}\bigg|_{\vec{a}'}\right) \nabla v_1(\vec{a}') + \left(\frac{\partial E_\alpha}{\partial v_2}\bigg|_{\vec{a}'}\right) \nabla v_2(\vec{a}') = 0$$

Then for some β, $-1 \leq \beta \leq 1$, $\nabla(\beta v_1 + (1-|\beta|)v_2) = 0$ at $\vec{a}'$. If $E_2\ (v_1,v_2)$ is concave at $v_1(\vec{a}')$, $v_2(\vec{a}')$ and $\beta \geq 0$, it can be shown that $\vec{a}'$ is a local minimum of $\beta v_1 + (1-\beta)v_2$.

Hence to minimize $E_\alpha(v_1(\vec{a}),v_2(\vec{a}))$ we need only consider functions of the form $\beta v_1(\vec{a}) + (1-|\beta|)v_2(\vec{a})$ for various values of β. These may be much easier to handle numerically. We now apply these results to optimal discrimination. Suppose we consider a certain class of discriminant functions $L(\vec{a})$. We may further restrict this class in such a way that $E(L(\vec{a})/\omega_2) - E(L(\vec{a})/\omega_1) = 1$. If the original parameter space A is rich enough to include the Gaussian discriminant and is homogeneous and translation invariant this may be easily achieved provided that the first and second moments are not identical under both hypotheses (see Theorem 1). If for the resulting set of the parameters the distribution functions of (L/ω_1) and of (L/ω_2) behave (asymptotically) as those of some family D of monotone error, we may determine that discriminant which (asymptotically) minimizes $E_{P(\omega_1)}(L/\omega_1, L/\omega_2)$ by considering the extrema of

$\beta \text{ Var } (L(\vec{a})/\omega_1) + (1-|\beta|) \text{ Var } (L(\vec{a})/\omega_2)$ for various values of β, $-1 \leq \beta \leq 1$, and either a) choosing that extremum which gives minimum error as calculated from knowledge of the family D or b) when D is not known, choosing that extremum for which the discriminant performs best on sample data.

One may at this point ask why we do not simply choose $L(\vec{a})$ which best separates the sample data. The reason in many practical situations is again that sample sizes are not exponential in d. If the parameter space A has dimension greater than or equal to d-1 (as is the case for the class of linear discriminants), it is highly likely that the "small" samples will be well separated by some $L(\vec{a})$. However this $L(\vec{a})$ will not necessarily perform well on new data since the $\vec{a}$ is not reliably estimated. Our procedure yields a <u>one</u> parameter family of discriminants determined from the (reliably estimated) moments. The sample size need not be exponential in d to reliably estimate the (single) parameter β.

In the following sections we will describe several applications of the above procedures and indicate which limit theorems in Probability guarantee the existence of the monotone error families D. Even if such limit theorems do not apply the procedure b) is indeed a second order solution *to the problem of finding the optimal discriminant* $L(\vec{a})$, in that an approximate solution is derived in terms of the means and variances of $L(\vec{a})$ under ω_1 and ω_2. We call this method the minimal variance solution.

IV. THE QUADRATIC DISCRIMINANT FUNCTION

By an affine transformation we may assume that $x_1, x_2, \ldots x_d$ are uncorrelated under both hypotheses and further that:

$$E(x_i/\omega_1) = 0; \; E(x_i/\omega_2) = 1$$

$$Var(x_i/\omega_1) = \lambda_i^1; \; Var(x_i/\omega_2) = \lambda_i^2$$

Let $Q(X) = \sum\limits_{i \leq j} a_{ij}\, x_i x_j + \sum b_i x_i$. In many cases Q is normal or (χ^2, n) for large d. The class of normals and the class of (χ^2, n) are classes which are characterized by their means and variances. It is easy to see that such classes are of monotone error. Hence the following minimal variance solution may indeed be (asymptotically) optimal:

$$E(Q/\omega_1) = \sum a_{ii} \lambda_i^1$$

$$E(Q/\omega_2) = \sum_{i<j} a_{ij} + \sum a_{ii}(1+\lambda_i^2) + \sum b_i$$

We want to find extrema of $\beta \operatorname{Var}(Q/\omega_1) + (1-|\beta|) \operatorname{Var}(Q/\omega_2)$ subject to $\sum\limits_{i<j} a_{ij} + \sum b_i + \sum a_{ii}\, (1+\lambda_i^2-\lambda_i^1) = 1$. Using a Lagrange multiplier Φ, we determine extrema of

$$\beta E_1 \Big(\sum_{i<j} a_{ij} x_i x_j + \sum a_{ii}\, (x_i^2 - \lambda_i^1) + \sum b_i x_i \Big)^2$$
$$+ (1-|\beta|) E_2 \Big(\sum_{i<j} a_{ij}(x_i x_j - 1) + \sum a_{ii}(x_i^2 - 1 - \lambda_i^2) + \sum b_i (x_i - 1) \Big)^2$$
$$- \Phi \Big(\sum_{i<j} a_{ij} + \sum b_i + \sum a_{ii}(1+\lambda_i^2-\lambda_i^1) - 1 \Big)$$

Differentiating with respect to a_{ij} and b_i and setting equal to zero we obtain:

$$2 \sum_{\ell<k} a_{\ell k} \Big[\beta E_1(x_\ell x_k x_i x_j) + (1-|\beta|) E_2(x_\ell x_k x_i x_j - 1) \Big]$$
$$+ 2 \sum b_\ell \Big[\beta E_1(x_\ell x_i x_j) + (1-|\beta|) E_2(x_\ell x_i x_j - 1) \Big]$$
$$+ 2 \sum a_{\ell\ell} \Big[\beta E_1(x_\ell^2 x_i x_j) + (1-|\beta|) E_2(x_\ell^2 x_i x_j - 1 - \lambda_\ell^2) \Big]$$
$$= \Phi \quad \text{for } i < j$$

$$
\begin{aligned}
2 \sum_{\ell<k} a_{\ell k} & \left[\beta E_1(x_\ell x_k x_i^2) + (1-|\beta|) E_2(x_\ell x_k x_i^2 - 1 - \lambda_i^2)\right] \\
& + 2 \sum b_\ell \left[\beta E_1(x_\ell x_i^2) + (1-|\beta|) E_2(x_\ell x_i^2 - 1 - \lambda_i^2)\right] \\
& + 2 \sum a_{\ell\ell} \left[\beta E_1(x_\ell^2 x_i^2 - \lambda_\ell^1 \ \lambda_i^1) + (1-|\beta|) E_2(x_\ell^2 x_i^2 - (1+\lambda_i^2)(1+\lambda_\ell^2)\right] \\
& = (1+\lambda_i^2 - \lambda_i^1) \ \Phi
\end{aligned}
$$

and

$$
\begin{aligned}
2 \sum_{\ell<k} a_{\ell k} & \left[\beta E_1(x_\ell x_k x_i) + (1-|\beta|) E_2(x_\ell x_k x_i - 1)\right] \\
& + 2 \sum a_{\ell\ell} \left[\beta E_1(x_\ell^2 x_i) + (1-|\beta|) E_2(x_\ell^2 x_i - 1 - \lambda_\ell^2)\right] \\
& + 2 \ b_i \left[\beta \ \lambda_i^1 + (1-|\beta|) \ \lambda_i^2\right] = \Phi
\end{aligned}
$$

In matrix form $M \vec{a} = \Phi \vec{d}$ yields $\vec{a} = \Phi \ M^{-1} \ \vec{d} = \Phi \ \vec{c}$. From the constraint $\Phi = (\sum_{i<j} c_{ij} + \sum c_i + \sum c_{ii} (1+\lambda_i^2 - \lambda_i^1) \ -1)^{-1}$ where c_{ij}, c_i, and c_{ii} are the components of $\vec{c}$ corresponding to a_{ij}, b_i, and a_{ii} of $\vec{a}$.

Unfortunately the above solution requires an inversion of a $\frac{d^2+3d}{2} \times \frac{d^2+3d}{2}$ matrix for each value of β. Also the asymptotic class may be unknown. The principal axes discriminant $P = \sum a_i \ x_i^2 + \sum b_i \ x_i$ has the advantages of being more numerically feasible (d equations in d unknowns) and behaving normally (large d, independent x_i) in many cases. Its minimal variance solution is as follows:

$$
\begin{aligned}
2 \sum b_\ell & \left[\beta E_1(x_\ell x_i^2) + (1-|\beta|) E_2(x_\ell x_i^2 \ - 1 - \lambda_i^2)\right] \\
+ 2 \sum a_\ell & \left[\beta E_1(x_\ell^2 x_i^2 \ - \lambda_\ell^1 \ \lambda_i^1) + (1-|\beta|) E_2(x_\ell^2 x_i^2 \ - \ (1+\lambda_\ell^2)(1+\lambda_i^2))\right] \\
= (1+\lambda_i^2 - \lambda_i^1) \ \Phi &
\end{aligned}
$$

$$2 \sum a_\ell \left[\beta E_1 (x_\ell^2 x_i) + (1-|\beta|) E_2 (x_\ell^2 x_i - 1 - \lambda_\ell^2)\right]$$
$$+ 2b_i \left[\beta \lambda_i^1 + (1-|\beta|) \lambda_i^2\right] = \Phi$$

The above system can be reduced to d equations in the d unknowns a_i and then solved for each value of β. The solution (a_i, b_i, Φ) for the constrained system is then obtained in a straight forward manner. Normal tables may be used to estimate the error for each β.

V. GENERAL DISCRIMINANT FUNCTIONS

We may use the procedure in III to determine higher order discriminants. However this may be numerically impractical even for discriminants of moderate order. Also for large dimensions d and relatively small ($<d^3$) sample sizes only marginal densities and their correlations may be reliably estimated. Hence we propose a method which depends only on marginal discriminants and their correlations. In two asymptotic situations upper bounds on the error rate are obtained.

Let us assume that $x_1, x_2, \ldots$ are uniformly bounded but relax any other previous assumption. (However in many practical cases "uncorrelating" the x_i may increase the applicability of a limit theorem.) Let $f_i(x_i)$ be a discriminant function for the one-dimensional random variable

x_i. For simplicity[2] let us assume that, for all discriminants $f_i(x_i)$ henceforth considered, $E(f_i/\omega_i) \neq E(f_i/\omega_2)$. One choice of f_i is (an estimate of) the log-likelihood ratio of the marginal densities of x_i, $\ln p_2^i(x_i) - \ln p_1^i(x_i)$. Consider the discriminant function

$$D(X) = \sum_1^d a_i f_i(x_i)$$

To obtain the minimal variance solution we find extrema of

$$\beta E_1(\Sigma \; a_i f_i(x_i) - \Sigma \; a_i E_1(f_i(x_i)))^2$$
$$+ \; (1-|\beta|) E_2(\Sigma \; a_i f_i(x_i) - \Sigma \; a_i E_2(f_i(x_i)))^2$$
$$- \; \Phi \; (\Sigma \; a_i E_2(f_i(x_i)) - \Sigma \; a_i E_1(f_i(x_i))$$

These are given by -

$$\vec{a} = \Phi \; M^{-1}\vec{b} = \Phi \; \vec{c}$$

where

$$b_i = E_2(f_i) - E_1(f_i)$$

$$m_{ij} = 2\beta E_1 \left[f_i f_j - E_1(f_i)E_1(f_j)\right]$$
$$+ \; 2(1-|\beta|)E_2 \left[f_i f_j - E_2(f_i)E_2(f_j)\right]$$

and $\Phi = \left(\Sigma \; c_i(E_2(f_i)-E_1(f_i))\right)^{-1}$

The following definition is adapted from [3].

Def. 3: An infinite sequence of random variables x_i is said to be finitely dependent if for every nonempty finite

[2] This is hardly a restriction since in practical cases estimates of the expectations mentioned will rarely be the same.

subset of the variables A there exists another finite subset B(A) (including A) such that $\{x_i \in A\}$ is independent of $\{x_i \in B^c\}$ and sup $\frac{|B(A)|}{|A|} < \infty$, where || denotes the cardinality of a set.

If our sequence $x_1, x_2, \ldots$ is finitely dependent then it follows that the sequence $\{f_i(x_i)\}$ is also finitely dependent. By the results of [3] the discriminant function $D(X) = \sum_1^d a_i f_i(x_i)$ will be normally distributed for large d provided that the a_i's become sufficiently small[3] as d becomes large. This has the following useful consequence.

Theorem 3: Let $x_1, x_2, \ldots$ be finitely dependent and suppose $f_i(x_i)$ is the log-likelihood ratio. Suppose further that $x_{i_1}, x_{i_2}, \ldots$ are independent. If for large d the minimal variance solution D(X) has coefficients which satisfy the above "smallness" conditions, then the probability of error using D(X) will become bounded above by the Bayes error for the sequence $x_{i_1}, x_{i_2}, \ldots, x_{i_k}$ $(i_k \leq d < i_{k+1})$.

outline of proof: the optimal discriminant function for

$x_{i_1}, x_{i_2}, \ldots x_{i_k}$ is $\overline{D}(X) = \frac{1}{k} \sum_1^k f_{i_s}(x_{i_s})$.

As d becomes large k becomes large and the above coefficients $(\frac{1}{k})$ will satisfy the "smallness"

[3] The exact conditions on the a_i in terms of the variances of the $f_i(x_i)$ can be determined from Corollary 4.2 on page 232 of [3] .

condition. Since the minimal variance solution is unique it will have minimum error among all those discriminants whose coefficients satisfy the "smallness" condition (since this is a connected set in the coefficient space).

In many practical situations the marginal densities are not available. We introduce a method of minimal marginal moment variance. (m.m.m.v.) Let x be a one-dimensional random variable. Suppose $y_1, y_2, \ldots$ is independent and identically distributed (i.i.d.) with the same distribution as x under both hypotheses. Consider a polynomial discriminant function $P_Q(Y) = \sum_1^d \sum_1^Q a_{ij} y_i^j$. The minimal variance solution will be of the form $P_Q(Y) = \frac{1}{d} \sum_1^d \sum_1^Q \bar{a}_j y_i^j$ where $\bar{a}_j$ are extrema of $\beta \operatorname{Var}(\sum_1^Q \bar{a}_j x^j/\omega_1) + (1-\beta) \operatorname{Var}(\sum_1^Q \bar{a}_j x^j/\omega_2)$

$$- \Phi \left(E_2(\sum_1^Q \bar{a}_j x^j) - E_1(\sum_1^Q \bar{a}_j x^j) - 1\right).$$

Differentiating yields -

$$2 \sum_1^Q \bar{a}_\ell \left[\beta E_1(x^{\ell+j} - x^j E_1(x^\ell)) + (1-\beta) E_2(x^{\ell+j} - x^j E_2(x^\ell))\right]$$

$$= \Phi \left(E_2(x^j) - E_1(x^j)\right).$$

Using normal tables the correct β is determined and the corresponding $\bar{a}_j$ calculated from the variances of P_Q.

Def. 4: The m.m.m.v. discriminant function of x of degree (d,Q) is given by $\bar{f}_Q(x) = \sum_1^Q \bar{a}_j x^j$ where the $\bar{a}_j$ are the above minimal variance coefficients.

Theorem 4: Let $x_1, x_2, \ldots$ be finitely dependent and suppose $f_i(x_i)$ is the m.m.m.v. discriminant function of x_i of degree (k,Q).

Suppose further that $x_{i_1}, x_{i_2}, \ldots$ are i.i.d. If for large d and large Q the minimal variance solution D(X) has coefficients which satisfy the previous "smallness" conditions, then the probability of error using D(X) will become bounded above by the Bayes error for the sequence $x_{i_1}, x_{i_2}, \ldots, x_{i_k}$ $(i_k \leq d < i_{k+1})$.

outline of proof: the optimal discriminant function $x_{i_1}, x_{i_2}, \ldots, x_{i_k}$ is $\overline{D}(X) = \frac{1}{k} \sum_1^k \ell(x_{i_s})$ where $\ell(x_{i_s})$ is the log likelihood ratio. Since for large Q this may be well approximated by $D_Q(X) = \frac{1}{k} \sum_1^k \sum_1^Q \overline{g}_j\ x_{i_s}^j$ where $\overline{g}_j$ are determined by fitting $\ell(x_{i_s})$ to a polynomial of degree Q. But for large k the minimal variance solution for $x_{i_1}, x_{i_2}, \ldots$ is $M(X) = \frac{1}{k}\sum_1^k \sum_1^Q \overline{a}_j\ x_{i_s}^j = \frac{1}{k}\sum_1^k \overline{f}_Q(x_{i_s})$. It will satisfy the smallness condition and its error will approach that of D_Q which in turn approaches that of $\overline{D}$. Hence the error of M approaches the Bayes error of $x_{i_1}, x_{i_2}, \ldots x_{i_k}$. By arguments similar to the proof of Theorem 3, the error of D(X) will become bounded by the Bayes error of $x_{i_1}, x_{i_2}, \ldots, x_{i_k}$.

We conclude from these theorems that, even if we do not know which subsequence $x_{i_1}, x_{i_2}, \ldots$ of the process $x_1, x_2, \ldots$ is an independent sequence, we may discriminate the finitely dependent process (asymptotically) as well as we could discriminate $x_{i_1}, x_{i_2}, \ldots$ if the $\{i_j\}$ were known.

REFERENCES

1. K. Fukunaga, *Introduction to Statistical Pattern Recognition,* (Academic Press, 1972).

2. C. W. Therrien, "A Sequential Approach to Target Discrimination," IEEE Trans. Aerospace and Electron. Systems AES-14, 433-440, (1978).

3. L. H. Y. Chen, "Two Central Limit Problems for Dependent Random Variables," Z. Wahrscheinlichkeitstheorie verw. Gebeite 43, 223-243 (1978).

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

19 **REPORT DOCUMENTATION PAGE**		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 18 ESD-TR-79-31	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* 6 On Optimal Discriminants Between Two Classes of Random Variables in Terms of the Moments of Their Distributions		5. TYPE OF REPORT & PERIOD COVERED 9 Technical Note 14 TN- 6. PERFORMING ORG. REPORT NUMBER Technical Note 1979-22
7. AUTHOR(s) 10 Lee K. Jones		8. CONTRACT OR GRANT NUMBER(s) 15 F19628-78-C-0002
9. PERFORMING ORGANIZATION NAME AND ADDRESS Lincoln Laboratory, M.I.T. P.O. Box 73 Lexington, MA 02173		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Program Element No. 63311F Project No. 627A 16
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Systems Command, USAF Andrews AFB Washington, DC 20331	11	12. REPORT DATE 28 February 1979 13. NUMBER OF PAGES 28
14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)* Electronic Systems Division Hanscom AFB Bedford, MA 01731	12 26p.	15. SECURITY CLASS. *(of this report)* Unclassified 15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

None

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

density   random variables   functions

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

For many problems of interest in statistical pattern recognition, density estimates for a random variable X of dimension d are unreliable unless the number of sample vectors is very large $(>2^d)$. For even moderately large d ~~(d > 12)~~, sample sizes are often insufficient, However, lower order moments ~~of the form $x_i^q x_j^r$~~ may be accurately estimated. In this paper we are concerned with the problem of optimally discriminating between two classes of random variables in terms of the available information about them of reasonable accuracy (their lower order moments). *In no case* do we make any assumption about the form of the probability densities of random variables X. (We do in some cases assume certain forms for the densities of functions of these random variables L(X).)

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

207650

1B